ACS OMEGA

Open Access

This article is licensed under CC-BY 4.0

http://pubs.acs.org/journal/acsodf

Article

Data-Driven Discovery of High-Performance EOR Surfactants: Coupling Molecular Dynamics with Active Learning

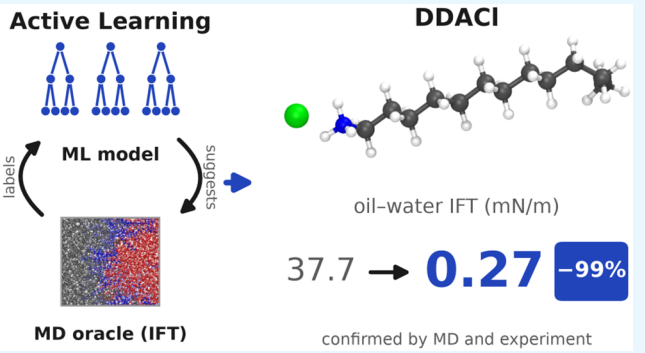
Eduardo Teixeira de Aguiar Veiga, Ednilsom Orestes, Renato D. Cunha, Paula Homem-de-Mello, Conny Cerai Ferreira, and James Moraes de Almeida*

Cite This: https://doi.org/10.1021/acsomega.6c10914

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Chemical Enhanced Oil Recovery (EOR) is a central route to maximizing hydrocarbon recovery from mature reservoirs. Although its field-scale success depends on several factors—including surfactant cost, adsorption and retention on the reservoir rock, solubility, and phase behavior—achieving a low oil–water interfacial tension (IFT) is a key enabling requirement. However, the design of surfactants meeting this requirement is impeded by the vast combinatorics of the chemical space and the prohibitive cost of traditional trial-and-error screening. In this study, we present a synergistic materials informatics framework that integrates Machine Learning (ML) and Molecular Dynamics (MD) simulations within an Active Learning (AL) workflow to accelerate surfactant discovery. Unlike static screening approaches, our methodology utilizes MD simulations as a high-fidelity oracle, iteratively labeling a subset of candidates to update a regressor, enabling the identification of high-performance surfactants from the PubChem database. Through this workflow, we identified Dodecylammonium Chloride (DDACl) as a top candidate; subsequent high-fidelity MD simulations confirmed a very low IFT at maximum surface coverage, outperforming widely used analogs such as Cetyltrimethylammonium Bromide (CTAB). We validated these computational findings through experimental tensiometry, confirming the efficacy of DDACl in reducing IFT in several oils with varying API gravities. The convergence of theoretical predictions with experimental results underscores the power of this data-driven framework to overcome the "small data" bottleneck in materials science, offering a scalable pathway for the data-driven discovery of surfactants with low oil–water IFT as candidates for EOR formulation development.



1. INTRODUCTION

The continued extraction of hydrocarbon resources remains a critical component of the global energy strategy during the current energy transition.[1] Despite advancements in primary and secondary recovery techniques, approximately 60% to 70% of the original oil in place (OOIP) often remains trapped within the reservoir rock matrix due to capillary forces and interfacial tension (IFT).[2,3] Chemical Enhanced Oil Recovery (EOR), particularly surfactant flooding, represents a vital technological frontier for mobilizing this residual oil. By reducing the oil–water IFT to ultralow values (10^{-3} mN/m) and altering rock wettability, surfactants can significantly increase the capillary number, facilitating the displacement of trapped hydrocarbons.[4,5] It is important to emphasize that a low interfacial tension (IFT) constitutes a necessary but not sufficient prerequisite for successful surfactant flooding. At the field scale, enhanced oil recovery (EOR) performance is additionally constrained by surfactant economics, adsorption and retention on reservoir minerals, aqueous solubility and tolerance to salinity, and the phase behavior of the coupled surfactant–oil–brine system.[6,7] Accordingly, the present study focuses specifically on the data-driven identification of surfactants that yield low oil–water IFT—i.e., the primary response variable targeted by our computational workflow—as an initial screening step, after which these complementary formulation and deployment criteria can be evaluated.

However, the design and selection of optimal surfactants constitute a high-dimensional optimization problem. The efficacy of a surfactant is governed by complex physicochemical interactions with crude oil, brine salinity, and reservoir temperature, effectively making the search space infinite.[8] Traditional discovery pipelines, characterized by trial-and-error synthesis and time-consuming laboratory screening, are increasingly viewed as insufficient to meet the economic and efficiency demands of the modern market.[9] Consequently, the

Received: September 14, 2026
Revised: October 19, 2026
Accepted: October 21, 2026

ACS Publications

© XXXX The Authors. Published by American Chemical Society

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

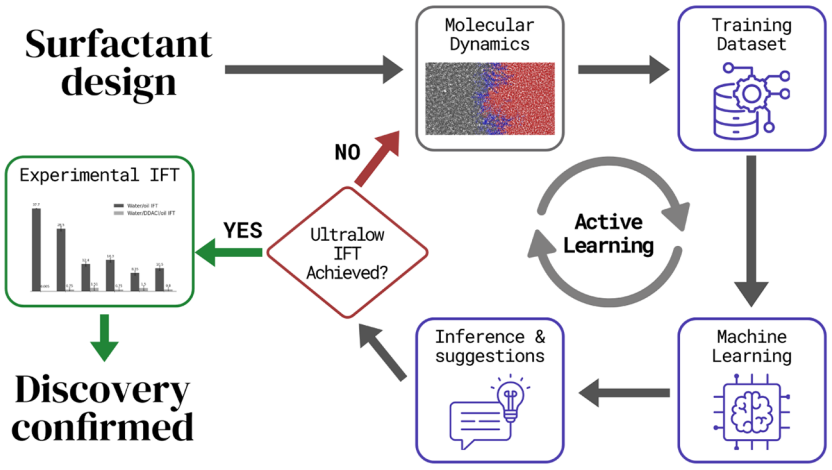


Figure 1. Methodology employed in this work is summarized in this figure. The process starts with molecular dynamics simulations.

field has witnessed a paradigm shift toward Materials Informatics, where Machine Learning (ML) and Molecular Dynamics (MD) are integrated to accelerate discovery.[10]

Recent literature highlights a bifurcation in computational discovery strategies. One dominant trend is the use of predictive ML models, such as tree-based ensembles[11,12] and Graph Neural Networks (GNNs)[13,14] to estimate interfacial tension (IFT) and critical micelle concentration (CMC) with high accuracy. The growing availability of curated surfactant data sets, such as SurfPro,[15] is beginning to ease the data scarcity that has long constrained such models. Building on this, emerging research has focused on *de novo* molecular generation. For example, Nnadili et al. (2024) introduced a framework combining Variational Autoencoders (VAEs) with Reinforcement Learning to generate novel surfactant structures.[16] Similarly, Sheikh et al. (2025) utilized diffusion models for the inverse design of nonionic surfactants,[17] while González-Núñez et al. (2025) applied multiobjective optimization to balance performance with sustainability constraints.[18] More recent graph-based generators, such as the GraphXForm graph transformer, further incorporate explicit structural constraints during molecular construction,[19] reflecting a broader maturation of generative approaches to molecular design.[20]

Although generative models offer the allure of exploring vast chemical spaces, they often face a "synthesizability gap"—the theoretically optimized molecules may be difficult, expensive, or impossible to synthesize on industrial scales.[21] To address this, a pragmatic alternative is the intelligent screening of existing chemical libraries. This approach, often implemented through Active Learning (AL), iteratively selects the most promising candidates from databases of commercially available compounds for high-fidelity validation.[22,23] Using MD simulations as a "computational oracle", AL strategies can navigate the trade-off between exploration (searching new chemical spaces) and exploitation (refining known high performers) with minimal computational cost.[24]

In this study, we propose a synergistic framework that couples Machine Learning with Molecular Dynamics under an Active Learning methodology to accelerate the discovery of surfactants with low oil−water IFT. Unlike purely generative approaches, our workflow screens readily available compounds extracted from the PubChem database, ensuring that the identified candidates are accessible for immediate experimental validation. We demonstrate the efficacy of this approach through the discovery of Dodecylammonium Chloride (DDACl), a cationic surfactant. Although bromide-based analogs like DTAB and CTAB are ubiquitous in the EOR literature,[25] DDACl has received comparatively less attention despite possessing unique interfacial properties. We validate the AL-driven prediction through MD simulations and experimental IFT measurements across crude oils of varying densities. These results confirm DDACl's ability to achieve very low IFT, underscoring the potential of data-driven workflows to identify promising IFT-reducing candidates for subsequent EOR formulation studies.

2. METHODOLOGY

Different computational methodologies and experimental analyses were used for the surfactant design. The objective is to find surfactants capable of significantly reducing the water/oil interfacial tension (IFT), preferably to ultralow values. The entire methodology is summarized in Figure 1. After the development and validation of a new computational oil model, the water/oil IFT values from molecular dynamics (MD) simulations for a few surfactants were used to train an ML model under the AL methodology to highlight new surfactants with better performance in IFT reduction. The performance of the surfactant suggested by the ML model was then confirmed by theoretical and experimental methods. In the following subsections, each step is described.

2.1. Computational Details

To calculate the IFT values of water/oil through MD simulations, a new computational oil model was developed according to the oil composition analysis conducted by França et al.,[26] which focused on oils from various basins within Brazilian presalt reservoirs. This oil model was validated with reference values for densities and IFT values from MD simulations for a variety of hydrocarbons individually, including alkanes, cycloalkanes, aromatics, and naphthenic acids, ranging from 3 to 45 carbon atoms. MD simulations to validate the oil model were performed using the Amber package,[27] employing the GAFF2[28] and Lipid14[29,30] force fields, along with the TIP3P[31] water model and 2 ns of thermalization in an NVT ensemble at 300 K using a Langevin thermostat[32] (γ = 1 ps^{-1}) with a time step of 2 fs (0.2 fs during heating). Covalent bonds involving hydrogen were constrained with the SHAKE algorithm, and long-range electrostatics were

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

handled with Particle Mesh Ewald (PME) using a real-space cutoff of 12.0 Å. Partial atomic charges were derived using the restrained electrostatic potential (RESP) method,[33] as implemented in the `antechamber` module of AmberTools,[27] fitted to the molecular electrostatic potential computed with the Gaussian code[34] at the B3LYP/6−31G* level using the Merz−Kollman scheme; this charge derivation protocol was applied to all organic molecules parametrized with GAFF2 in this work. Initial configurations were built by packing the molecules into the simulation box with Packmol,[35] using a 2.0 Å tolerance to prevent atomic overlaps, and the corresponding topologies were assembled with the LEaP module of AMBER. The density was allowed to equilibrate for 2.5 ns at an NPT ensemble using a Berendsen weak-coupling barostat[36] (isotropic scaling, τ_P = 0.5 ps), at 300 K and 1 atm, followed by a production run of 5 ns, where the interfacial tension was sampled for every step of the simulation. The initial simulated system sizes were 100 × 100 × 250 Å^3, as shown in Figure 2, and the IFT values were obtained with the following equation:[37−40]

$$\gamma = \frac{1}{2} L_Z \left[P_{zz} - \frac{1}{2}(P_{xx} + P_{yy}) \right] \quad (1)$$

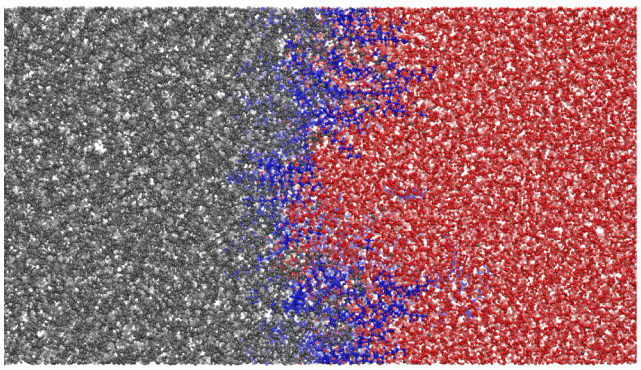


Figure 2. Image of the interfacial region of a simulated system. The red atoms are water, blue atoms the surfactant, and gray atoms the oil.

It is important to note that the 5 ns production runs used throughout the AL screening cycles are intentionally short. Their purpose is to provide a reliable *relative ranking* of surfactants to guide phase-space exploration, not to deliver converged absolute IFT values. Active learning in materials and molecular science routinely operates on approximate oracle labels: convergence of the surrogate model is what matters for efficient candidate selection, not the numerical precision of each label.[41,42] Once DDACl was identified as the top-performing candidate, an independent high-fidelity simulation was performed, whose protocol is described in Section 2.2.

The data set of surfactant features was obtained from the PubChem API database and the chemparse library.[43] For each surfactant, features based on type and molecular composition were selected: surface excess concentration, Molecular weight, topological polar surface area, complexity, charge, hydrogen bond donor count, hydrogen bond acceptor count, rotational bond count, heavy atom count, isotope atom count, atomic stereo count, defined atom stereo count, undefined atom stereo count, bond stereo count, defined bonded bond stereo count, undefined bond stereo count, covalent unit count, carbon atom count, hydrogen atom count, bromine atom count, nitrogen atom count, sodium atom count, oxygen atom count, sulfur atom count, fluorine atom count, chlorine atom count, phosphorus atom count, and iodine atom count.

The ML model was based on Python code with the Scikit-Learn library,[44] and four different regression algorithms were considered in this work: support vector machines, k-nearest neighbors, gradient boosting, and random forest.[11,12,37,45] To select the most suitable regression algorithm for the ML model, the K-Fold Cross-Validation technique was used. The initial data set for the first active learning run had 16 surfactants, drawn from PubChem to span a broad range of charge types (cationic, anionic, zwitterionic, and nonionic) and molecular weights, providing initial coverage of the surfactant feature space. It is a small data set, but active learning does not need to provide a precise regression; it just needs to help sample the phase space. Thus, starting with a small data set is usual. Learning from limited labeled data is a well-recognized challenge in molecular and materials science, and active learning is among the strategies developed to address it.[41,42] This initial data set was divided into K = 4 partitions, with K-1 partitions used to train the ML model with each regression algorithm, while the remaining one was used to validate them. The process ended when all partitions were used as a validation set.[46] Feature vectors were standardized (zero mean, unit variance) before training the SVR and k-NN algorithms; tree-based methods (gradient boosting and random forest) were trained on the raw features, as scaling does not affect their performance. The generalization of the selected regression algorithm was performed with the Leave-One-Out Cross-Validation technique, a K-Fold Cross-Validation in which the number of partitions equals the size of the data set (K = 16).

The search for suitable surfactants was performed using the active learning (AL) methodology. In this methodology, the target property is optimized through cycles of training/prediction of the ML model, with some strategic data from the predicted set being labeled by an external source of information and added to the training set of the next cycle. In our case, the external source of information was the simulated IFT with MD. A data set of water/oil IFTs for 16 surfactants was used in the first optimization cycle of the AL methodology. From the first predicted set of surfactants, four were chosen to have their water/oil IFT values determined by MD simulation and subsequently added to the training set for the next (second) optimization cycle. The next three cycles proceeded similarly until the fifth optimization cycle, in which a surfactant with a substantially low water/oil IFT value was detected. The AL process was stopped when a surfactant achieving ultralow IFT (i.e., greater than 99.9% reduction relative to the bare water/oil IFT of 37.70 mN/m) was identified, and no candidate in the current predicted set was forecast by the ML model to outperform it.

2.2. DDACl High-Fidelity Simulation

Following the identification of DDACl as the top candidate by the AL workflow, a dedicated large-scale simulation was performed to obtain a converged IFT under conditions corresponding to maximum surface coverage. The surface excess concentration for this simulation was set to Γ_{max} = 1.63 μmol m^{-2} (minimum molecular area A_{min} = 101.8 Å^2 molecule^{-1}) and was experimentally determined from the Gibbs adsorption isotherm (eq 2) applied to the pendant drop IFT measurements described in Section 2.3. For a 100 × 100 Å

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

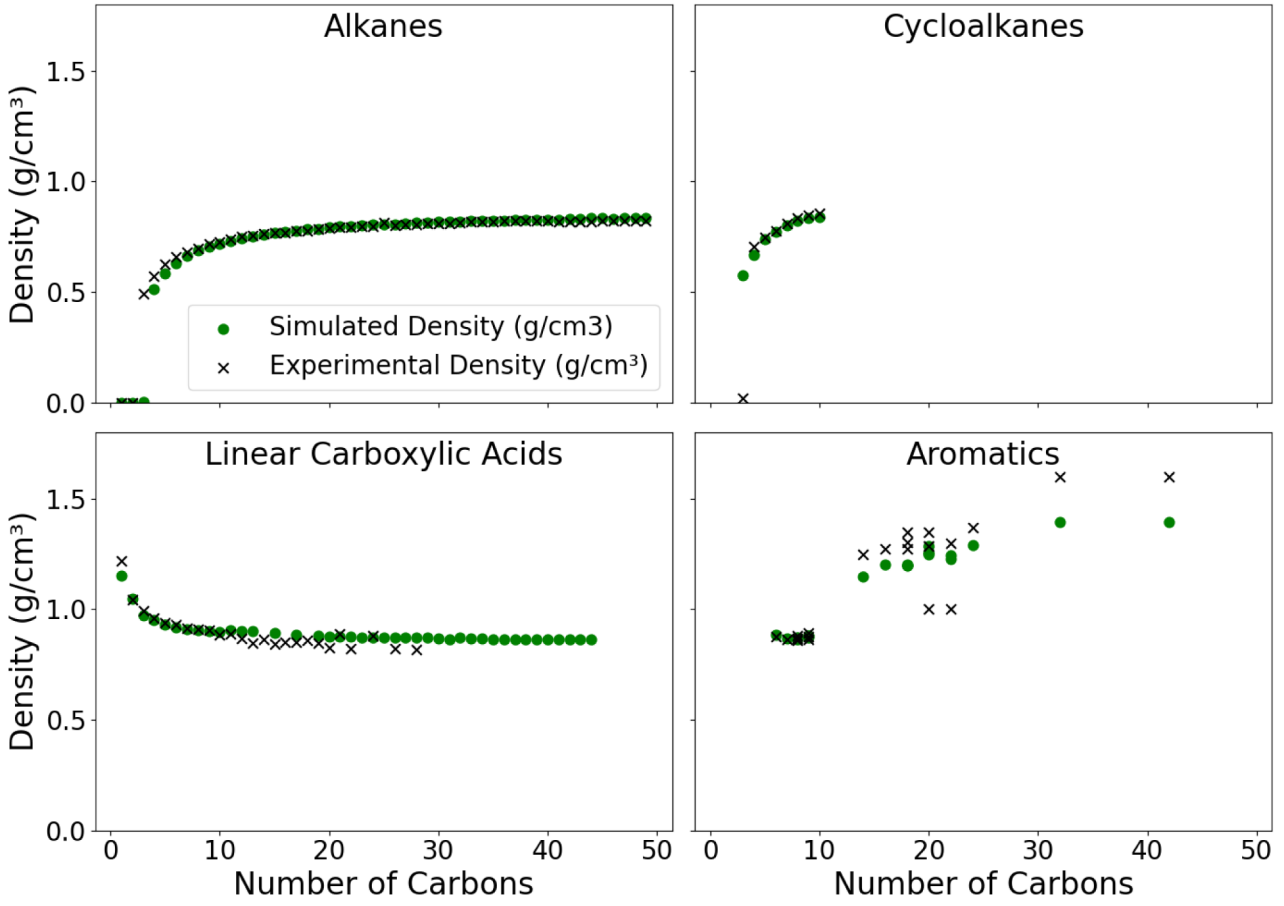


Figure 3. Densities obtained with the simulation compared to experimental values.[51−54]

simulation box, this corresponds to 98 DDACl molecules per leaflet (196 total for the two interfaces present in the periodic slab).

$$\Gamma_{max} = -\frac{1}{nRT}\frac{d\gamma}{d\ln C} \quad (2)$$

where n = 2 for a 1:1 ionic surfactant (DDACl → DDA^+ + Cl^-) in the absence of added salt, R is the universal gas constant, T the temperature, and $d\gamma/d\ln C$ the slope of the IFT−concentration curve in the steepest pre-CMC linear regime.

All MD simulations were performed with the AMBER 24 software package[47] using the GPU-accelerated `pmemd.cuda` module. The simulation system represents a DDACl-stabilized oil−water bilayer slab with initial box dimensions of 100 × 100 × 224 Å; after pressure equilibration, the box relaxed to approximately 97.5 × 97.5 × 214.3 Å. The complete system contained 37,220 molecules and 194,787 atoms, parametrized with GAFF2 and TIP3P, with atomic charges derived from B3LYP/6−31G* as described above. The equilibration protocol comprised energy minimization (100,000 steps, steepest descent followed by conjugate gradient), NVT heating from 3 K to 300 K over 2 ps (Δt = 0.2 fs, Langevin thermostat), a further 200 ps NVT run at 300 K (Δt = 2 fs), and a 200 ns NPT equilibration (Berendsen barostat, P = 1.01325 bar, Langevin thermostat).

Production trajectories were collected under quasi-NVT conditions: the Berendsen barostat was retained with a pressure-relaxation time of τ_P = 100,000 ps, effectively freezing box dimensions on the 100 ns production time scale while keeping the virial-computation pathway active for GPU-accelerated pressure-tensor output. Temperature was maintained at 300 K with the Andersen thermostat.[48] Six independent replicas of 100 ns each (Δt = 2 fs) were launched from the same equilibrated checkpoint but with distinct initial velocity distributions (random seeds), giving a total of 600 ns of IFT data. The instantaneous IFT was written at every MD step, giving $N = 5 \times 10^7$ samples per replica (Δt_s = 0.002 ps). The IFT was computed from eq 1 using Welford's online algorithm for the running mean within each replica, with block averaging (block sizes 0.1−20 ns) to estimate within-replica uncertainty. The grand mean and inter-replica standard error were computed across the six replica means.[49]

2.3. Experimental Methods

The water/air surface tension and the water/oil interfacial tension were both measured with a pendant-drop tensiometer (Teclis Instruments, France) at room temperature (20°C). The critical micelle concentration (CMC) and water/air surface tension were determined with DDACl/water solutions prepared in the following concentration range: 0.00, 0.001, 0.005, 0.010, 0.050, 0.100, 0.250, 0.500, 0.750, 1.000, 1.500, 2.000%m m^{-1}. An earlier determination of these quantities using the Wilhelmy plate method in a force tensiometer (Force Tensiometer K100, Krüss), together with interfacial tension measurements of DDACl against additional crude oils of different API gravity, is reported in the Supporting Information.

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

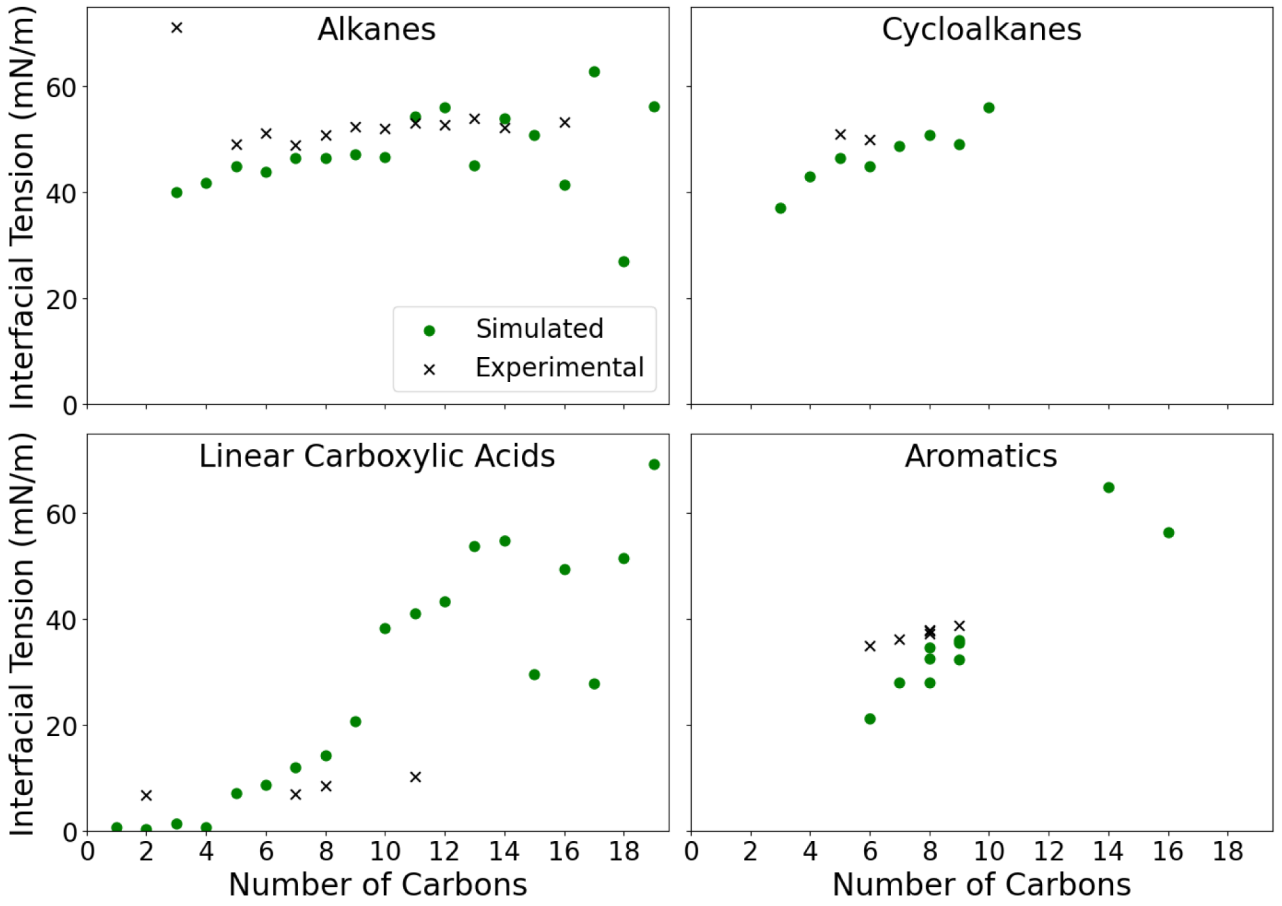


Figure 4. Comparison between the simulated interfacial tensions and experimental values.[55−59]

The tensiometer is equipped with an optical image acquisition system and a camera for monitoring drop geometry, and it uses Drop Shape Analysis software for data acquisition and processing; the drop profile was continuously recorded during each experiment and analyzed by fitting the Young−Laplace equation to determine the interfacial properties, with an instrument precision of ± 0.1 mN/m. For the water/oil measurements, an oil drop with an approximate volume of 8 μL was formed at the end of a U-shaped needle using a 500 μL syringe and subsequently inserted into a cell containing 20 mL of the aqueous phase, and the evolution of the oil−water interface was monitored over time.

DDACl/water solutions at concentrations of 0.001, 0.010, 0.050, 0.100, 0.500, and 1.000%m/m were tested against mineral oil. The maximum surface excess concentration Γ_{max} and minimum molecular area A_{min} were determined by fitting eq 2 to the steepest pre-CMC linear regime of the γ vs ln C curve (0.01−0.10% m/m, R^2 = 0.992).

Alternatively to density (specific gravity), the *API* degree (eq 3) can be used to classify crude oil as heavy, medium, or light.

$$\text{API} = \frac{141.5}{SG} - 131.5 \tag{3}$$

in which the specific gravity, SG, is the ratio between the oil density, ρ, and the water density, ρ_0: $SG = (\rho/\rho_0)$.

In this scale, the API of heavy crude oils ranges from 10.0° to 22.3°, medium oils from 22.3° to 31.1°, and light oils from 31.1° and higher. The mineral oil considered in this work has an API degree of 34.80°.

3. RESULTS AND DISCUSSION

3.1. Oil Model

Many different molecules, including alkanes, cycloalkanes, aromatics, and carboxylic acids, can comprise crude oil. The decision on which molecules of each class will take part in the computational oil model depends first on the oil that is trying to be modeled, but also on how the MD force field performs on each molecule individually.

Figure 3 presents the experimental and MD-simulated densities of alkanes, cycloalkanes, aromatics, and linear carboxylic acids over a wide range of carbon numbers for each class. Alkanes, the larger class, show excellent agreement between the experimental and MD-simulated densities, consistent with the good performance reported for GAFF in systematic benchmarks of organic-liquid densities.[50] The exception is propane, whose normal boiling point of 231 K[51] lies well below the 300 K simulation temperature, so it is not a liquid under the simulated conditions. Linear carboxylic acid experimental densities are also satisfactorily predicted by MD simulations, with small deviations for the highest densities in the trend. Cyclic compounds consist of two smaller classes: cycloalkanes and aromatics. For cycloalkanes, the simulated MD densities agree well with the experimental data, except for cyclopropane, which, for the same reason, is not a liquid at 300 K; its normal boiling point being 240 K.[51] The experimental densities of aromatics with less than 10 carbons are also fairly reproduced by the MD simulations; however, for larger molecules, the force field shows low shifted densities compared to the experimental ones.

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

Table 1. Composition of the Oil Used in the Simulations (Mass Fractions), with Average Molecular Weight of 205.34 g/mol and an API Grade of 36.95[a]

Name	MW(g/mol)	Conc. (%)	Composition by type (%)		Composition by compounds (%)	
Nonanoic acid	158.24	10.23	22.00	Acid	50.56	Paraffinic
Octanoic acid	144.21	7.45				
Hexanoic acid	116.16	2.32				
Pentanoic acid	102.13	2.00				
Tetracontane	563.10	10.00	28.56	Alkane		
Hexatriacontane	507.00	6.30				
Tricosane	324.60	4.16				
Hexadecane	226.44	3.07				
Decane	142.28	3.03				
Octane	114.23	2.00				
1,3,5-Trimethylbenzene	120.19	9.05	22.00	Aromatic	49.44	Naphthenic
Paraxylene	106.16	7.05				
Toluene	92.14	3.90				
Benzene	78.11	2.00				
Cyclodecane	140.27	11.18	27.44	Cycloalkanes		
Cyclononane	126.24	8.58				
Cyclooctane	112.21	5.68				
Cycloheptane	98.19	2.00				

[a]MW stands for molecular weight and Conc. for concentration.

The IFT values between water and alkanes, cycloalkanes, aromatics, and linear carboxylic acids were also obtained with MD simulations and compared with experimental data (Figure 4). Even with the scarcity of experimental data, the simulated/experimental results are in good agreement, with only one larger deviation, again for propane. The trend of increasing IFT with carbon number is well described for the four classes. However, for alkanes with more than 20 carbons and for cycloalkanes, aromatics, and linear acids with more than 9 carbons, the melting point exceeds 300 K, indicating solid compounds at this temperature. Although such properties can be addressed by simulated annealing and larger simulation boxes, our focus remained on creating a computational oil model capable of representing various types of hydrocarbons while maintaining a liquid state at 300 K. This would happen if the oil composition is not dominated by molecules with high melting points.

Once the MD force field applicability was established for different types of oil components individually, an oil model composition could be designed. The computational oil model developed in this work aligns with the SAN-07[26] type of oil from the Santos Basin, which has an estimated density of 0.879 g cm^{-3} and an API gravity of 29.4. On the scale above, SAN-07 falls in the medium range and is composed of 56% saturated hydrocarbons, 22% aromatic hydrocarbons, and 22% polar hydrocarbons.[26] Moreover, it has a notably low concentration of nitrogen-containing compounds and a prevalence of oxygen-containing derivatives, primarily carboxylic acids.[26]

Table 1 lists the components of our oil model with their respective mass-fraction contributions to the composition. As SAN-07,[26] our oil model is composed of 56% saturated hydrocarbons, 22% aromatic hydrocarbons, and 22% polar hydrocarbons, although our model lacks nitrogen-containing compounds. The molecular mass of the oil model is 205.34 g·mol^{-1} with a simulated density of 0.84 g·cm^{-3} and an API gravity of 36.95. The API value of the model (36.95) is higher than that of SAN-07 (29.4). Two practical constraints prevent a closer match. First, the largest hydrocarbons of a real crude could not be packed and equilibrated satisfactorily within a simulation cell of computationally feasible size, so the heaviest fractions are under-represented; this lowers the average molecular weight of the model and consequently raises its API. Second, asphaltenes, which strongly affect the density and API of real crudes, were not included at this stage. The model, therefore, reproduces the class-level composition of SAN-07 (saturated/aromatic/polar fractions) but should be regarded as inspired by SAN-07 rather than as a faithful reproduction of it. Long-chain *n*-alkanes such as tetracontane and hexatriacontane are common constituents of crude oils. Although they are wax-like as pure compounds at 300 K, they do not render the mixture waxy as long as they are not among its major components: they remain dissolved in the liquid matrix formed by the more abundant lighter hydrocarbons, exactly as they do in real oils. Consistently, the oil phase remained a single homogeneous liquid throughout the simulations, with no signs of crystallization or phase separation, and its simulated density (0.84 g·cm^{-3}) is typical of a liquid light crude, which would not be the case if any component had solidified. The oil−water simulation box contained 3,588 oil molecules in contact with 33,400 water molecules (16,700 in each of the two water regions); the number of molecules of each oil component is listed in the Supporting Information. The water/oil IFT value without surfactant obtained with the MD simulation was 37.70 *mN/m*, which will serve as a reference point for simulations including surfactants.

3.2. Regression Algorithm

Table 2 shows the SEC values for 16 randomly chosen surfactants extracted from the PubChem database and the chemparse library, together with their respective water/oil IFT values obtained from MD simulations as described in the Methodology section. For each surfactant, the number of molecules placed at the two oil−water interfaces of the box follows directly from its SEC (in molecules nm^{-2}) and the 100 Å × 100 Å (= 100 nm^2) area of each interface.

The data set from Table 2 was first used to select the regression algorithm of the ML model with the K-Fold Cross-Validation technique. Figure 5 compares the IFT values of

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

Table 2. Initial Dataset of 16 Surfactants with Their Respective Pubchem Compound ID (CID), Name, Surface Excess Concentration (SEC) in (#/nm^2), and Simulated Water/Oil IFT Values in (mN/m)

CID	Name	SEC (molecules/nm^2)	Simulated IFT (mN/m)
5974	CTAB	1.48	4.16
3423265	SDS	3.19	17.00
443315	Polysorbate 80n(PS 80)	3.00	92.98
5590	Triton X-100	2.10	22.43
15099	beta-Alanine	2.24	27.21
23677967	SNS	1.83	14.88
23662403	SDBS	1.53	6.07
2735067	SL	2.20	21.18
11046239	SML	2.48	8.37
1752	4-NP	2.41	31.21
247	BET	2.67	36.60
1088	SAR	1.09	34.95
84703	DDAPS	1.25	24.03
156694001	P-CAB agent 1	2.86	56.07
43043	LABA	2.41	21.51
456370	Srr SB3	1.85	23.24

water/oil from MD simulations with those predicted by the ML model using each of the four regression algorithms. The water/oil IFT values of all 16 surfactants predicted by the SVR algorithm range from 20 to 25 mN/m, clearly differing from the simulated IFT values (Figure 5 a). Similarly, the k-NN algorithm also predicted water/oil IFT values in disagreement with the simulated ones, although slightly more dispersive, ranging from 15.0 to 35.0 mN/m (Figure 5 b). Both the SVR and k-NN algorithms behave closely to a trivial average guess, thus overrating low IFT values and underrating high ones. In contrast, the gradient boosting algorithm provides water/oil IFT values that match exactly with all 16 simulated IFT values (every point lies on top of the ideal line), which usually happens in overfitted models (Figure 5 c). However, the best reliable result of the four regression algorithms tested in this work was provided by the random forest algorithm (Figure 5 d), which also shows a tendency to overrate low IFT values of water/oil and underrate high ones like SVR and k-NN, in a much less drastic manner. The random forest results follow a trend line close to the ideal relationship, mainly for the average IFT values of water/oil in Table 2, and are therefore a potential candidate to compose the ML model. Thus, the most suitable model for continuing with the suggestions for the active learning process is the random forest regressor. This preference for tree-based ensembles is consistent with previous studies that benchmarked regression algorithms for IFT prediction, where random forest and related methods provided the most reliable estimates[11,12] here, however, the labels are obtained from MD simulations rather than empirical correlations, and the regressor is used to guide phase-space sampling rather than to produce precise IFT values.

3.3. Surfactant Discovery

With the selected regression algorithm, the active learning (AL) methodology was adopted to suggest new surfactants to be simulated. The data set in Table 2 was used as the training set for the first optimization cycle in the AL methodology. Then, an inference was performed on 17 different surfactants to predict their IFT, as shown in Table 3.

Four surfactants from the first predicted set (Table 3) were strategically selected for water/oil IFT determination with MD simulations, which play the role of an external source of information for the active learning process. From these predicted IFT values, we selected the highest, the lowest, and two with medium IFT values. This strategy is intended to help explore the phase space rather than just exploiting it for the lowest IFT. With this, we hope to reduce model error at each iteration, yielding better suggestions for each AL step until we find a suitable surfactant.

One can see in Table 3 that the simulated IFT is significantly different from the model prediction, although we still simulated two with low IFT and two with high IFT, which should contribute to decreasing the model error for the low- and high-IFT regions of the phase space. Subsequently, these simulated IFT were added to the training set for a second optimization cycle in the AL methodology (Table S1 - Supporting Information). The second predicted set was then generated by the ML model (Table S2 - Supporting Information), and from this, new surfactants were selected to label with MD simulations. These selected surfactants were then added to the training set for another cycle to continue the optimization process. Five optimization cycles were performed in total, and Tables S1 to S6 show the training and predicted sets from the second to fifth cycles.

The strategy of wider exploration of different regions of the IFT phase space is intended when surfactants are labeled with MD simulations in each predicted set, rather than choosing only those surfactants with the lowest IFT values. The aim is to develop an ML model capable of recognizing trends rather than providing accurate IFT numerical results. These trends can reveal which surfactants are likely suitable or unsuitable for our oil model. A suitable surfactant is one with a predicted IFT value significantly below the pure water/oil IFT value (no surfactant) of 37.7 mN/m, while an unsuitable one has a predicted IFT value close to or above this reference point.

After five iterations, the ML model discovered surfactants with exceptional reduction capabilities of the water/oil IFT. Table 4 shows the ten surfactants with the lowest water/oil IFT values, according to MD simulations, which the ML model suggested in the five AL optimization cycles. From Table 4, four surfactants can reduce the IFT of water/oil by 80−90%, namely CPC, DTAB, SDBS, and CTAB. Another five surfactants can reduce the IFT in water/oil by more than 90%, namely TTAB, C14mimBr, C16mimBr, OTAB, and DDAB, which represented an almost 99% reduction in IFT, from 37.70 to 0.53 mN/m. Finally, DDACl presented the best screening performance of all surfactants considered by reducing the water/oil IFT by 99.98%, reaching a value of 0.0048 mN/m in the short screening simulation (see footnote[a] in Table 4). The converged IFT obtained from the high-fidelity replica simulation is reported in Section 3.4.

It is important to note that 0.0048 mN/m is the result of the MD screening simulation, not a machine learning prediction. The ML model guided the selection of DDACl as a top candidate; the IFT value was independently determined by MD.

Some of these surfactants have been extensively studied in the literature, such as CTAB[40,60−63] and less frequently its relatives DTAB, TTAB, and OTAB, in addition to SDS and the imidazolium ionic liquids C14mimBr and C16mimBr. In particular, the bromide anion (Br^{-1}) is a striking feature of the best surfactants, as seven out of ten have this feature, except for

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

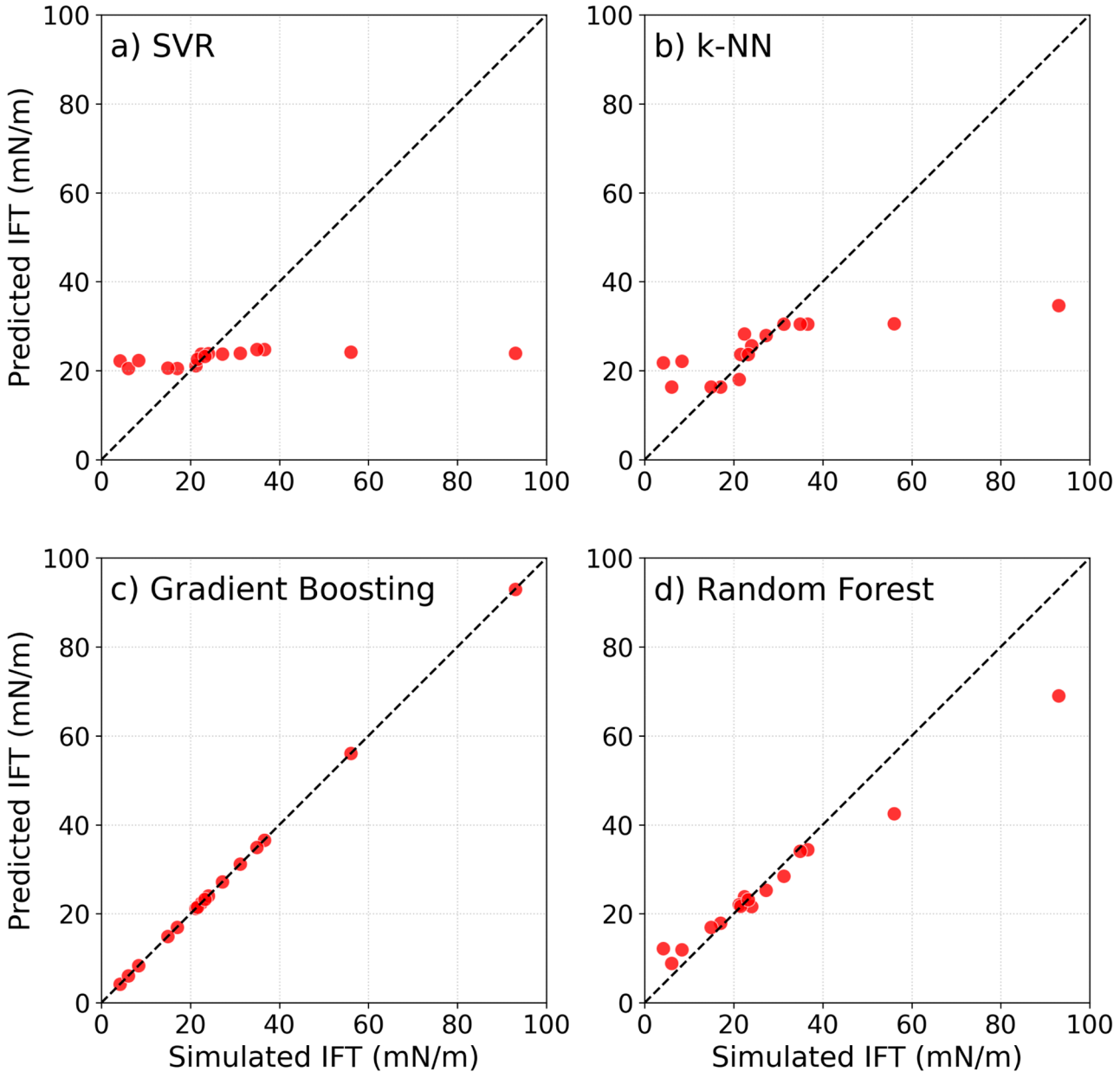


Figure 5. Values of the predicted IFT compared to simulated values for different regression techniques.

DDACl, which is highlighted as the most effective surface agent at the water/oil interface and carries a chlorine anion Cl^{-1}. Although some studies focusing on properties such as viscosity[64] and surface tension (water/air)[65] of DDACl could be found, the same is not true for applications in EOR, for which no published study has been detected. In addition to the mined SEC value for DDACl of 1.34 surfactant/nm^2, a lower concentration of 1.21 surfactant/nm^2 was also simulated, and the IFT value of water/oil obtained was 0.396 mN/m, which is higher than the previous 0.0048 mN/m, but still a low interfacial tension.

3.4. DDACl High-Fidelity Simulation Results

The interfacial tension of the DDACl-laden oil−water interface at the maximum surface-excess concentration Γ_{max} was computed from six independent 100 ns production replicas (600 ns of interfacial tension data). Table 5 summarizes the mean IFT and the within-replica block-averaged standard error for each replica.

The instantaneous IFT exhibits large thermal fluctuations with a standard deviation $\sigma \approx 79.8$ mN m^{-1}, characteristic of pressure-tensor-based IFT calculations in finite-area simulation cells.[66] The naïve standard error under full statistical independence is $SE_{naive} = \sigma/\sqrt{N} \approx 0.011$ mN m^{-1} (with $N = 5 \times 10^7$ samples per replica at $\Delta t_s = 0.002$ ps). The within-replica block-averaged standard error (~0.21−0.31 mN m^{-1}) substantially exceeds this value, confirming strong temporal correlations in the IFT time series. The integrated autocorrelation time, back-calculated from the block-SE plateau as $\tau_{int} \approx (g - 1)/2\ \Delta t_s$ with $g = (SE_{block}/SE_{naive})^2$, is approximately 0.35−0.75 ps per replica (a lower bound, as the block-SE curve had not fully plateaued at the largest block size tested), yielding a statistical inefficiency $g \approx 350$−750 and an effective sample count $N_{eff} = N/g \approx 7 \times 10^4$−$1.5 \times 10^5$. The

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

Table 3. Table with All Predicted Interfacial Tension (IFT) Values by the Machine Learning Model, for the First Run of the Active Learning Process, as Well as the Simulated Values for the Chosen Surfactants

	IFT (mN/m)	
Surfactant CID	Predicted	Simulated
77520435	12.55	1.34
71340974	17.07	
10428412	20.20	
14250	23.13	3.58
5353376	23.89	33.95
14249	24.60	
2726	25.79	
135398640	25.86	
88069	29.32	
6285	30.80	
2733501	32.651	
124081846	32.90	
12219	34.00	
5917	35.09	
11015971	36.76	
90472707	43.46	
104772	45.28	30.91

Table 4. IFT Reduction in Comparison to the Water/Oil System (No Surfactant), Whose Simulated IFT is 37.70 mN/m

CID	Name	IFT simulated (mN/m)	IFT reduction (%)	Opt cycle
31239	CPC	6.8776	81.75	third
14249	DTAB	6.7280	82.15	second
23662403	SDBS	6.0668	83.90	Training
5974	CTAB	4.1607	88.96	Training
14250	TTAB	3.5762	90.51	first
71340974	C16mimBr	3.2580	91.35	second
77520435	C14mimBr	1.3370	96.45	first
70708	OTAB	1.1395	96.97	fifth
second	NDEAB	0.5322	98.58	fourth
139078665	DDAB	0.5322	98.58	fourth
13582	DDACl	0.0048[a]	99.98	fifth

[a]Screening estimate from the short (5 ns) simulation protocol used to rank candidates in the AL workflow. The converged IFT from the dedicated high-fidelity replica simulation is reported in Section 3.4.

Table 5. Mean Interfacial Tension and Block-Averaged Standard Error for Each Independent Production Replica

Replica	$\bar{\gamma}$ (mN m^{-1})	Block SE (mN m^{-1})
1	0.405	0.308
2	0.014	0.307
3	0.244	0.293
4	0.551	0.296
5	0.139	0.212
6	0.268	0.222
Grand mean	0.270	0.078[a]

[a]Inter-replica standard error.

corresponding corrected standard error $\sigma/\sqrt{N_{\text{eff}}} \approx 0.21-0.31$ mN m^{-1} is self-consistent with the block estimates and indicates the presence of slow collective motions of the surfactant layer on time scales exceeding the fast thermal decorrelation.[67] Running independent replicas effectively averages over these slow modes: the inter-replica standard error of 0.078 mN m^{-1} is approximately three to four times smaller than the within-replica block estimate, confirming that six replicas provide well-converged statistics.

The grand mean IFT across all six replicas is

$$\bar{\gamma} = 0.270 \pm 0.078 \text{ mNm}^{-1} \text{ (95\%CI:[0.118, 0.422] mNm}^{-1}\text{)} \quad (4)$$

where the uncertainty is the inter-replica standard error and the confidence interval is computed as $\bar{\gamma} \pm 1.96$ SE. Empirical force fields can marginally overstabilize tightly packed surfactant monolayers and thereby slightly underestimate the IFT.[68−70] The small residual value obtained here should be interpreted in that light. Physically, the result indicates that DDACl at Γ_{max} reduces the oil−water IFT to a few tenths of mN m^{-1}, more than 2 orders of magnitude below the bare oil−water value of 37.70 mN m^{-1}.

3.5. Experimental Results

Figure 6 shows how the surface tension (water/air) of the water varies with the concentration of surfactant. It can be seen that the surface tension of the water drops steeply to values around 27.0 mN/m for a very small increase in DDACl concentration. Above this small concentration, the surfactant molecules begin to aggregate and no longer affect the surface tension, which then remains nearly constant. The intersection point of the trendlines (red) for both regimes determines the CMC of DDACl, which is equal to 0.047% m/m with a water surface tension of 27.8 mN/m. This is lower than those found for other cationic surfactants, such as DTAB with 29.3 mN/m, CTAB with 35.1 mN/m, BTAB with 47.5 mN/m, and TBAB with 49.0 mN/m,[25] thus demonstrating good performance.

Figure 7 shows the pendant-drop IFT of the DDACl/water−mineral oil system as a function of DDACl concentration. A steep decrease in IFT is observed in the pre-CMC region, followed by a plateau near the CMC. Fitting the Gibbs adsorption isotherm (eq 2) to the linear regime (0.01−0.10% m/m, $R^2 = 0.992$) yields $\Gamma_{max} = 1.63$ μmol m^{-2} and $A_{min} = 101.8$ Å^2 molecule^{-1}, which were used to set the surface concentration of DDACl in the high-fidelity MD simulation (Section 2.2). At 1.0%m/m DDACl—above the CMC and corresponding to Γ_{max} conditions—the measured IFT is 0.746 ± 0.140 mN/m, a reduction of nearly 2 orders of magnitude relative to the surfactant-free interface. This experimental result is in qualitative agreement with the simulation grand mean of 0.270 ± 0.078 mN/m; both indicate that DDACl at maximum packing reduces the oil−water IFT to a few tenths of mN/m. The remaining quantitative discrepancy of approximately 0.5 mN/m can be related to the mineral oil not being exactly the same as the digital oil model, and it is consistent with the expected accuracy of empirical force fields for surfactant systems as well as the sensitivity of very low IFT values to small errors in the intermolecular potential.[68−70]

The reduction of the water/oil interfacial tension by DDACl was further evaluated against a set of crude oils spanning a range of API gravities. These measurements, performed with a Wilhelmy plate force tensiometer, are reported in the Supporting Information and show a consistent reduction of the water/oil IFT to values around 1 mN/m across all tested oils at 1.0%m/m DDACl, corroborating the very low interfacial tensions found for the mineral oil and computational systems.

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

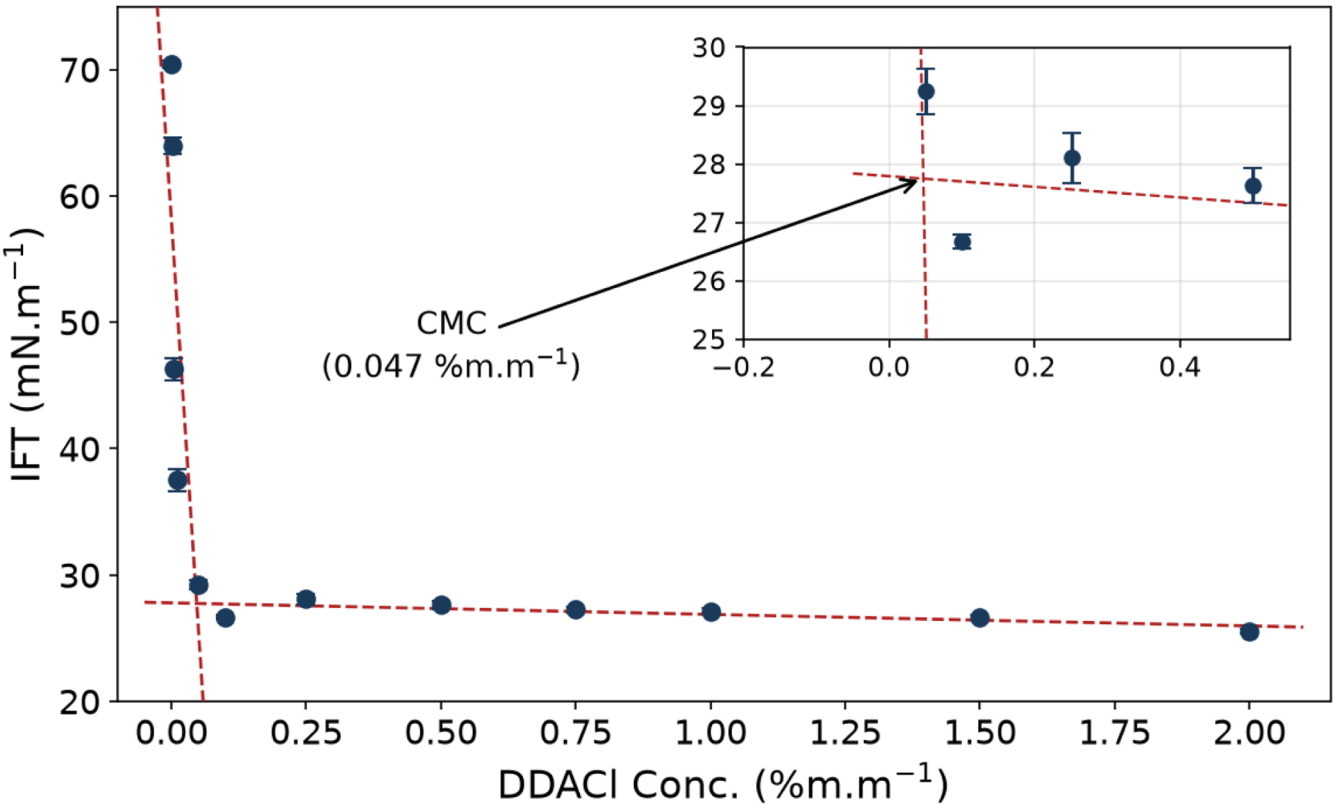


Figure 6. Experimental determination of the DDACl critical micellar concentration at the water/air interface.

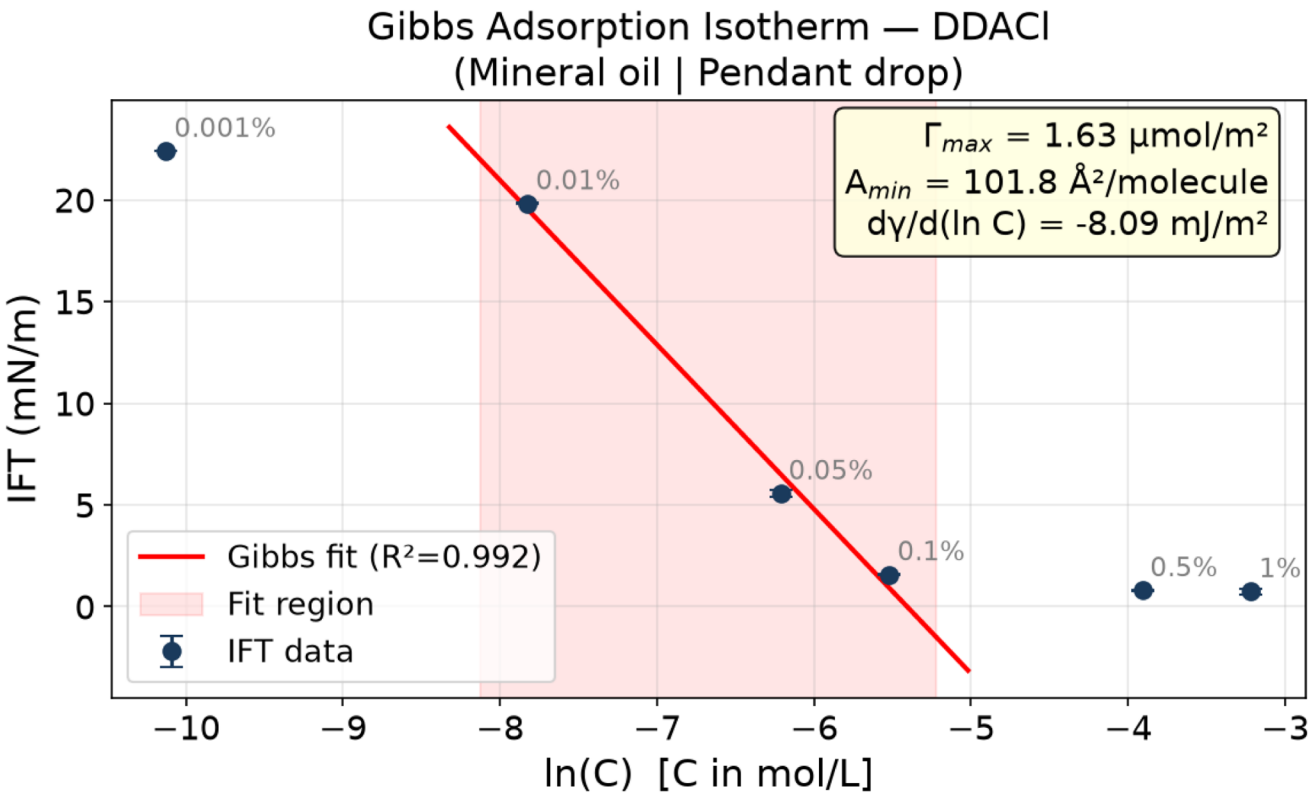


Figure 7. Pendant-drop IFT of the DDACl/water−mineral oil system as a function of DDACl concentration, with the Gibbs adsorption isotherm linear fit (dashed line) used to determine Γ_{max} and A_{min}.

It is instructive to interpret the reported sub-1 mN/m interfacial tension (IFT) values in the context of pore-scale oil mobilization, which is commonly parametrized by the capillary number, $N_c = v\mu/\gamma$, where v denotes the Darcy velocity, μ

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

denotes the viscosity of the displacing fluid, and γ denotes the IFT. Capillary desaturation measurements in sandstones indicate that residual oil saturation decreases appreciably only after N_c exceeds a critical threshold of approximately 10^{-5}, and that substantial desaturation typically requires an additional increase of roughly 2 orders of magnitude in N_c.[71] At representative flooding velocities, these conditions correspond to the ultralow-IFT regime (10^{-2}–10^{-3} mN/m) targeted in classical low-tension flooding approaches.[72]

Within this framework, lowering the IFT from the baseline oil–water value of 37.70 mN/m to the 0.27–0.75 mN/m range reported here increases N_c by approximately 2 orders of magnitude, whereas further reductions within the sub-1 mN/m interval (e.g., from 1 to 0.2 mN/m) yield less than one additional order-of-magnitude increase. Consequently, additional gains in oil mobilization generally require access to the ultralow-IFT regime, which in field applications is achieved not by the neat surfactant alone but through formulation strategies, including optimization of salinity, the use of cosurfactants and cosolvents, and control of middle-phase microemulsion behavior.[6,72] Accordingly, a surfactant such as DDACl, which reduces IFT intrinsically to a few tenths of mN/m at low concentration and without formulation aids, is best viewed as a promising starting point for formulation development rather than a stand-alone enhanced oil recovery (EOR) solution.

4. CONCLUSIONS

The selection of the most suitable surfactants for enhanced oil recovery (EOR) through chemical flooding is a daunting task. In this study, Machine Learning (ML) models and Molecular Dynamics (MD) simulations were combined under the Active Learning (AL) methodology to predict the best-performing surfactant in reducing the water/oil interfacial tension (IFT) values, a key property for EOR. The performance of the surfactant highlighted by the AL process was theoretically and experimentally verified for oil samples with different API values.

To evaluate the impact of surfactants on the water/oil interfacial tension (IFT) using molecular dynamics (MD) simulations, a new computational crude oil model with an API value of 37.0, based on the composition of a presalt crude oil from the Santos Basin, was developed and validated. The validation process was performed against the densities and IFT values with water for all individual oil compound molecules, improving the oil model and allowing new ones to be designed.

The most suitable regression algorithm for the ML model, the random forest, was selected by comparing the water/oil IFT values predicted by ML and simulated by MD for 16 surfactants. Proper selection and generalization of the regression algorithm are crucial procedures to avoid results close to trivial average guesses, as well as overfitting, thus ensuring more reliable predictions of the ML model on unseen data.

The choice of Active Learning as the methodology to obtain predictions from the Machine Learning model for surfactants capable of significantly reducing water/oil IFT values was important since label acquisition can be costly (MD simulation of all surfactants), and rapid improvement of the ML model with minimal annotation effort is desired. The short (5 ns) MD simulations used as the AL oracle intentionally prioritize relative ranking over numerical precision; they serve to map the surfactant phase space efficiently rather than to deliver converged absolute IFT values. Also important was the proper selection of new surfactants from the predicted set of each run to be labeled with MD simulations and added to the training set of the next run, i.e., surfactants with low, medium, and high predicted values of water/oil IFT. The strategy of exploring a wider region of the chemical space of features by looking for suitable surfactants rather than exploiting it with an optimization focused on a specific property seems to be determinant in achieving, after just five runs of the AL methodology, the identification of DDACl as a top candidate.

The exceptional IFT-reduction capability of DDACl was confirmed by a dedicated high-fidelity simulation (six independent replicas × 100 ns = 600 ns of interfacial-tension data) at the experimentally determined maximum surface excess concentration (Γ_{max} = 1.63 μmol m^{-2}), yielding a converged IFT of 0.270 ± 0.078 mN/m—a very low value, more than 2 orders of magnitude below the bare oil–water IFT. This was experimentally confirmed by pendant-drop tensiometry, which measured an IFT of 0.746 ± 0.140 mN/m at 1.0%m/m DDACl in mineral oil, as well as values around 1 mN/m against additional crude oils of varying API gravity (Supporting Information). Furthermore, the CMC for the water/air surface tension was determined to be 0.047%m m^{-1}, suggesting that DDACl could be effective at low concentration.

This research showcases the promising coupling of molecular dynamics simulations with machine learning to optimize surfactant discovery. The proposed methodology constitutes a valuable framework for the design and selection of surfactants and provides directions for experiments that can validate the findings.

Finally, it is worth mentioning that the methodology developed in this work can be adapted to numerous applications, as well as to study other perspectives on smart surfactant discovery, such as IFT for different oil models and environmental conditions, while accounting for enhanced oil recovery and fully or easily recycling surfactants. Because attaining a low oil–water interfacial tension (IFT) constitutes only one prerequisite for an effective surfactant-flooding process, a logical extension of the present framework is to incorporate additional screening criteria—namely surfactant cost; adsorption and retention on reservoir minerals; aqueous solubility and salinity tolerance; and surfactant–oil–brine phase behavior.[6,7] Such an expansion would transform the current single-objective workflow into a multiobjective discovery platform aimed at identifying field-deployable formulations.

■ ASSOCIATED CONTENT

Data Availability Statement

All the data used for training and the inferences are available in the Supporting Information as csv files. Furthermore, a PDF file is provided with tables of the data and captions with a brief explanation.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.6c10914.

Tables with all interactions for machine learning (PDF)

All data used for training and the inferences in csv files (ZIP)

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX–XXX

AUTHOR INFORMATION

Corresponding Author

James Moraes de Almeida − *Centro de Ciências Naturais e Humanas (CCNH) − Universidade Federal do ABC, Santo André, São Paulo 09020-580, Brazil; Ilum School of Science, Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, São Paulo 13083-970, Brazil;* orcid.org/0000-0002-3126-7619; Email: james.almeida@ufabc.edu.br

Authors

Eduardo Teixeira de Aguiar Veiga − *Centro de Ciências Naturais e Humanas (CCNH) − Universidade Federal do ABC, Santo André, São Paulo 09020-580, Brazil*

Ednilsom Orestes − *Departamento de Ciências Exatas, Escola de Engenharia Industrial Metalúrgica de Volta Redonda, Universidade Federal Fluminense, Volta Redonda, Rio de Janeiro 27255-125, Brasil*

Renato D. Cunha − *Centro de Ciências Naturais e Humanas (CCNH) − Universidade Federal do ABC, Santo André, São Paulo 09020-580, Brazil*

Paula Homem-de-Mello − *Centro de Ciências Naturais e Humanas (CCNH) − Universidade Federal do ABC, Santo André, São Paulo 09020-580, Brazil;* orcid.org/0000-0002-7049-4689

Conny Cerai Ferreira − *Escola de Engenharia Industrial Metalúrgica de Volta Redonda, Universidade Federal Fluminense, Volta Redonda 27255-125, Brazil*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.6c10914

Funding

The Article Processing Charge for the publication of this research was funded by the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES), Brazil (ROR identifier: 00x0ma614).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank FAPESP (2021/13554-0, 2023/09820-2, 2025/09430-5), CAPES, CNPq (308403/2025-8), and the INCT Materials Informatics for their financial support. The authors would also like to thank CCM-UFABC and the National Laboratory for Scientific Computing (LNCC/MCTI, Brazil) for providing computational resources.

REFERENCES

(1) Precedence Research. *Enhanced Oil Recovery Market Size, Share, and Trends 2025 to 2034. Market Research Report 2025*; 2025. https://www.precedence-research.com/.

(2) Muggeridge, A.; Cockin, A.; Webb, K.; Frampton, H.; Collins, I.; Moulds, T.; Salino, P. Recovery rates, enhanced oil recovery and technological limits. *Philos. Trans. R. Soc., A* **2014**, *372*, 20120320.

(3) Fact.MR. *Chemical Enhanced Oil Recovery Market Outlook (2025−2035). Market Research Report*; 2025.

(4) Massarweh, O.; Abushaikha, A. S. The use of surfactants in enhanced oil recovery: A review of recent advances. *Energy Rep.* **2020**, *6*, 3150−3178.

(5) Chowdhury, S.; Shrivastava, S.; Kakati, A.; Sangwai, J. S. Comprehensive review on the role of surfactants in the chemical enhanced oil recovery process. *Ind. Eng. Chem. Res.* **2022**, *61*, 21−64.

(6) Hirasaki, G. J.; Miller, C. A.; Puerto, M. Recent advances in surfactant EOR. *Spe J.* **2011**, *16*, 889−907.

(7) Kamal, M. S.; Hussein, I. A.; Sultan, A. S. Review on surfactant flooding: Phase behavior, retention, interfacial tension, and field applications. *Energy Fuels* **2017**, *31*, 7701−7720.

(8) Rosen, M. J.; Kunjappu, J. T. *Surfactants and interfacial phenomena*; John Wiley & Sons, 2012. DOI: .

(9) Takahashi, K.; Takahashi, L. Toward the golden age of materials informatics: perspective and opportunities. *J. Phys. Chem. Lett.* **2023**, *14*, 4726−4733.

(10) Zuccarini, C.; Ramachandran, K.; Jayaseelan, D. Material discovery and modeling acceleration via machine learning. *APL Mater.* **2024**, *12*, 090601.

(11) Rashidi-Khaniabadi, A.; Rashidi-Khaniabadi, E.; Amiri-Ramsheh, B.; Mohammadi, M.-R.; Hemmati-Sarapardeh, A. Modeling interfacial tension of surfactant−hydrocarbon systems using robust tree-based machine learning algorithms. *Sci. Rep.* **2023**, *13*, 10836.

(12) Yousefmarzi, F.; Haratian, A.; Mahdavi Kalatehno, J.; Keihani Kamal, M. Machine learning approaches for estimating interfacial tension between oil/gas and oil/water systems: a performance analysis. *Sci. Rep.* **2024**, *14*, 858.

(13) Brozos, C.; Rittig, J. G.; Bhattacharya, S.; Akanny, E.; Kohlmann, C.; Mitsos, A. Graph neural networks for surfactant multi-property prediction. *Colloids Surf., A* **2024**, *694*, 134133.

(14) Brozos, C.; Rittig, J. G.; Bhattacharya, S.; Akanny, E.; Kohlmann, C.; Mitsos, A. Predicting the Temperature Dependence of Surfactant CMCs Using Graph Neural Networks. *J. Chem. Theory Comput.* **2024**, *20*, 5695−5707.

(15) Hödl, S.; Hermans, L.; Dankloff, P.; Piruska, A.; Huck, W. T. S.; Robinson, W. E. SurfPro − a curated database and predictive model of experimental properties of surfactants. *Digital Discovery* **2025**, *4*, 1176−1187.

(16) Nnadili, M.; Okafor, A. N.; Olayiwola, T.; Akinpelu, D.; Kumar, R.; Romagnoli, J. A. Surfactant-Specific AI-Driven Molecular Design: Integrating Generative Models, Predictive Modeling, and Reinforcement Learning for Tailored Surfactant Synthesis. *Ind. Eng. Chem. Res.* **2024**, *63*, 6313−6324.

(17) Sheikh, S. I.; Navya, V. S.; Sharma, R.; Roy, S.; Singh, J. K.Digital Surfactant*arXiv***2025**

(18) González-Núñez, S.; Sumer, Z.; Amador, C.; Madhav, P.; Muralidharan, A.; Adjiman, C. S.; Martín, M. Design of Surfactant Molecules Under Performance Constraints. *ACS Sustainable Chem. Eng.* **2025**, *13*, 13808−13824.

(19) Pirnay, J.; Rittig, J. G.; Wolf, A.; Grohe, M.; Burger, J.; Mitsos, A.; Grimm, D. G. GraphXForm: graph transformer for computer-aided molecular design. *Digital Discovery* **2025**, *4*, 1052−1065.

(20) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1608.

(21) Guo, J.; Schwaller, P. Directly optimizing for synthesizability in generative molecular design using retrosynthesis models. *Chem. Sci.* **2025**, *16*, 6943−6956.

(22) Lookman, T.; Balachandran, P. V.; Xue, D.; Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *Npj Comput. Mater.* **2019**, *5*, 21.

(23) Wang, Y.; Samarasinghe, D. S. N. D.; Deng, H.; Liu, C.; Aikens, C. M. Gaussian Process Approach to Constructing Transferable Force Fields for Thiolate-Protected Gold Nanoclusters. *J. Chem. Inf. Model.* **2025**, *65*, 3892−3902.

(24) Jung, G. S.; Choi, J. Y.; Lee, S. M. Active learning of neural network potentials for rare events. *Digital Discovery* **2024**, *3*, 514−527.

(25) Kumar, S.; Panigrahi, P.; Saw, R. K.; Mandal, A. Interfacial interaction of cationic surfactants and its effect on wettability alteration of oil-wet carbonate rock. *Energy Fuels* **2016**, *30*, 2846−2857.

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

(26) França, D.; Coutinho, D. M.; Barra, T. A.; Xavier, R. S.; Azevedo, D. A. Molecular-level characterization of Brazilian pre-salt crude oils by advanced analytical techniques. *Fuel* **2021**, *293*, 120474.
(27) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 198−210.
(28) He, X.; Man, V. H.; Yang, W.; Lee, T.-S.; Wang, J. A fast and high-quality charge model for the next generation general AMBER force field. *J. Chem. Phys.* **2020**, *153*, 114502.
(29) Dickson, C. J.; Madej, B. D.; Skjevik, A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C. Lipid14: the amber lipid force field. *J. Chem. Theory Comput.* **2014**, *10*, 865−879.
(30) Madej, B. D.; Gould, I. R.; Walker, R. C. A parameterization of cholesterol for mixed lipid bilayer simulation within the Amber Lipid14 force field. *J. Phys. Chem. B* **2015**, *119*, 12424−12435.
(31) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(32) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N'-Methylamide. *Biopolymers* **1992**, *32*, 523−535.
(33) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.
(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A., et al. *Gaussian 09, Revision D. 01*; 2009, Gaussian, Inc.: Wallingford CT.
(35) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009**, *30*, 2157−2164.
(36) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(37) Kirch, A.; Celaschi, Y. M.; de Almeida, J. M.; Miranda, C. R. Brine−oil interfacial tension modeling: Assessment of machine learning techniques combined with molecular dynamics. *ACS Appl. Mater. Interfaces* **2020**, *12*, 15837−15843.
(38) de Almeida, J. M.; Miranda, C. R. Improved oil recovery in nanopores: NanoIOR. *Sci. Rep.* **2016**, *6*, 28128.
(39) de Almeida, J. M.; Miranda, C. R. Confinement and hydrophilicity effects on geologically relevant fluids in silica nanopores. *Phys. Rev. Fluids* **2020**, *5*, 083801.
(40) de Almeida, J. M.; Ferreira, C. C.; Bandeira, L.; Cunha, R. D.; Coutinho-Neto, M. D.; Homem-De-Mello, P.; Orestes, E.; Nascimento, R. S. V. Synergistic interaction of hyperbranched polyglycerols and cetyltrimethylammonium bromide for oil/water interfacial tension reduction: A Molecular dynamics study. *J. Phys. Chem. B* **2023**, *127*, 9356−9365.
(41) Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine Learning Methods for Small Data Challenges in Molecular Science. *Chem. Rev.* **2023**, *123*, 8736−8780.
(42) Xu, P.; Ji, X.; Li, M.; Lu, W. Small data machine learning in materials science. *Npj Comput. Mater.* **2023**, *9*, 42.
(43) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51*, D1373−D1380.
(44) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(45) Omrani, S.; Ghasemi, M.; Singh, M.; Mahmoodpour, S.; Zhou, T.; Babaei, M.; Niasar, V. Interfacial tension−temperature−pressure−salinity relationship for the hydrogen−brine system under reservoir conditions: integration of molecular dynamics and machine learning. *Langmuir* **2023**, *39*, 12680−12691.
(46) Jung, Y. Multiple predicting K-fold cross-validation for model selection. *J. Nonparamet. Stat.* **2018**, *30*, 197−215.
(47) Case, D. A.; et al. Recent Developments in Amber Biomolecular Simulations. *J. Chem. Inf. Model.* **2025**, *65*, 7835−7843.
(48) Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72*, 2384−2393.
(49) Grossfield, A.; Patrone, P. N.; Roe, D. R.; Schultz, A. J.; Siderius, D. W.; Zuckerman, D. M. Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **2019**, *1*, 5067.
(50) Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant. *J. Chem. Theory Comput.* **2012**, *8*, 61−74.
(51) Haynes, W. M. *CRC handbook of chemistry and physics*; CRC press: Boca Raton, 2016. DOI: .
(52) Setzmann, U.; Wagner, W. A new equation of state and tables of thermodynamic properties for methane covering the range from the melting line to 625 K at pressures up to 1000 MPa. *J. Phys. Chem. Ref. Data* **1991**, *20*, 1061−1155.
(53) O'neil, M. J. *The Merck index: an encyclopedia of chemicals, drugs, and biologicals*; RSC Publishing, 2013.
(54) Lide, D. R. *Handbook of Data on Organic Compounds*; Springer, 1989; Vol. 7.
(55) Wiegand, G.; Franck, E. Interfacial tension between water and non-polar fluids up to 473 K and 2800 bar. *Ber. Bunsengesellschaft Phys. Chem.* **1994**, *98*, 809−817.
(56) Demond, A. H.; Lindner, A. S. Estimation of interfacial tension between organic liquids and water. *Environ. Sci. Technol.* **1993**, *27*, 2318−2331.
(57) Zhang, S.; Lu, X.; Wu, J.; Tong, W.; Lei, Q.; Fang, W. Interfacial tensions for system of n-heptane+ water with quaternary ammonium surfactants and additives of NaCl or C2−C4 alcohols. *J. Chem. Eng. Data* **2014**, *59*, 860−868.
(58) Goebel, A.; Lunkenheimer, K. Interfacial tension of the water/n-alkane interface. *Langmuir* **1997**, *13*, 369−372.
(59) Aman, Z. M.; Zerpa, L. E.; Grasso, G.; Leith, W. E.; Sloan, E. D.; Sum, A. K.; Koh, C. A.Interfacial tension and mineral adhesion properties of cyclopentane hydrate*SPE/AAPG/SEG Unconventional Resources Technology Conference*Denver, Colorado, USA2013
(60) Chen, L.; Zhang, G.; Wang, L.; Wu, W.; Ge, J. Zeta potential of limestone in a large range of salinity. *Colloids Surf., A* **2014**, *450*, 1−8.
(61) Ferreira, C. C.; da Silva, T. B. G.; Francisco, A. D. D. S.; Bandeira, L.; Cunha, R. D.; Coutinho-Neto, M. D.; Homem-de Mello, P.; Moraes de Almeida, J.; Orestes, E.; Nascimento, R. S. V. Hyperbranched polyglycerols derivatives as cetyltrimethylammonium bromide nanocarriers on enhanced oil recovery processes. *J. Appl. Polym. Sci.* **2022**, *139* (9), 51725.
(62) Pan, F.; Zhang, Z.; Zhang, X.; Davarpanah, A. Impact of anionic and cationic surfactants interfacial tension on the oil recovery enhancement. *Powder Technol.* **2020**, *373*, 93−98.
(63) Matsubara, H.; Nakano, T.; Matsuda, T.; Takiue, T.; Aratono, M. Effects of Alkyl Chain Length on Synergetic Adsorption and Micelle Formation in Homologous Cationic Surfactant Mixtures. *Langmuir* **2005**, *21*, 8131−8137.
(64) Kamranfar, P.; Jamialahmadi, M. Effect of surfactant micelle shape transition on the microemulsion viscosity and its application in enhanced oil recovery processes. *J. Mol. Liq.* **2014**, *198*, 286−291.
(65) Motomura, K.; Iwanaga, S.-I.; Hayami, Y.; Uryu, S.; Matuura, R. Thermodynamic studies on adsorption at interfaces: IV. Dodecylammonium chloride at water/air interface. *J. Colloid Interface Sci.* **1981**, *80*, 32−38.
(66) Gloor, G. J.; Jackson, G.; Blas, F. J.; de Miguel, E. Test-Area Simulation Method for the Direct Determination of the Interfacial Tension of Systems with Continuous or Discontinuous Potentials. *J. Chem. Phys.* **2005**, *123*, 134703.

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX

(67) Nickerson, S.; Frost, D. S.; Phelan, H.; Dai, L. L. Comparison of the Capillary Wave Method and Pressure Tensor Route for Calculation of Interfacial Tension in Molecular Dynamics Simulations. *J. Comput. Chem.* **2013**, *34*, 2707−2715.
(68) Yu, L.; Gao, S.; Ding, W.; Bao, X.; Wang, H.; Yuan, R. Mechanism Analysis and Property Prediction of Extended Surfactants Based on the Respectively Optimized Force Field. *Langmuir* **2023**, *39*, 14859−14868.
(69) Li, J.; Amador, C.; Wilson, M. R. Computational Predictions of Interfacial Tension, Surface Tension, and Surfactant Adsorption Isotherms. *Phys. Chem. Chem. Phys.* **2024**, *26*, 12107−12120.
(70) Jójárt, B.; Martinek, T. A. Performance of the General Amber Force Field in Modeling Aqueous POPC Membrane Bilayers. *J. Comput. Chem.* **2007**, *28*, 2051−2058.
(71) Chatzis, I.; Morrow, N. Correlation of capillary number relationships for sandstone. *Soc. Pet. Eng. J.* **1984**, *24*, 555−562.
(72) Foster, W. A low-tension waterflooding process. *J. Pet. Technol.* **1973**, *25*, 205−210.

Downloaded from pubs.acs.org/acsodf/article-pdf/doi/10.1021/acsomega.6c10914/6911518/acsomega.6c10914.pdf by guest on 10 November 2026

https://doi.org/10.1021/acsomega.6c10914
ACS Omega XXXX, XXX, XXX−XXX